La visualización en la recuperación de información: estándares, tendencias y limitaciones

Mario Pérez-Montoro

14 octubre 2013

Pérez-Montoro, Mario (2014). "La visualización en la recuperación de información: estándares, tendencias y limitaciones". *Anuario ThinkEPI*, v. 8, pp. 301-306.

Resumen: La evolución de la disciplina de la recuperación de información no ha sido sensible al desarrollo de la visualización de resultados, ignorando que en muchas ocasiones la presentación de esos resultados juega un papel tan importante en la satisfacción de la necesidad de información del usuario como una buena selección de documentos del fondo. Es posible identificar una serie de modelos estándar y ciertas tendencias en la presentación visual de resultados fruto de la interrogación de un sistema. Se analizan esos modelos y tendencias y se establecen una serie de requisitos (arquitectónicos y semánticos) que pueden servir de guía para la mejora de la funcionalidad de las herramientas de visualización en el proceso de la recuperación.

Palabras clave: Visualización de la información, Recuperación de información, Modelos de visualización, Presentación visual, Página de resultados, Ecuación de búsqueda, Lenguaje de interrogación.

Title: Visualization in information retrieval: Standards, trends and limitations

Abstract: The evolution of the information retrieval discipline has not been sensitive to developing the display of results. In many cases, the results presentation is as important as the selection of the retrieved documents in satisfying the user's information needs. However, it is possible to identify a number of standard models and certain trends in the visual presentation of results in a retrieval system. This study analyzed the patterns and trends and established a series of requirements (architectural and semantic) as a guide for improving the functionality of visualization tools in the retrieval processes.

Keywords: Information visualization, Information retrieval, Visualization models, Visual presentation, Results page, Query search, Query language.

1. Introducción

La recuperación de información es una estrategia, basada en la interrogación, para la rápida localización de documentos de un fondo que puedan satisfacer las necesidades informativas de un usuario.

Según el modelo clásico, el proceso de la recuperación de información mediado por un sistema se estructura sobre tres pilares básicos (**Bates**, 1989):

- la necesidad de información del usuario (un estado mental): representada en el sistema mediante una ecuación de búsqueda perteneciente a un lenguaje de interrogación;
- el documento: que se somete a un proceso de representación de su contenido semántico;
- el mapeo o comparación entre la representación de la información contenida en el documento y la ecuación de búsqueda para identificar qué documentos pueden satisfacer la necesidad informativa del usuario.

Los documentos seleccionados tras el mapeo entre las dos representaciones (documentos-necesidad) son ofrecidos al usuario a través de una página de resultados que permite acceder a los mismos.

Tradicionalmente en la disciplina de la recuperación, el grueso de esfuerzos económicos e intelectuales se han invertido en el desarrollo y mejora de algoritmos cada vez más eficaces para la representación documental y el mapeo (**Baeza-Yates**; **Ribeiro-Neto**, 2011), descuidándose en muchas ocasiones la investigación en la presentación visual de los resultados.

La evolución de la disciplina no ha sido sensible al desarrollo de la visualización ignorando que la presentación de resultados juega un papel

tan importante como la buena selección de documentos. Una mala o no adecuada presentación puede dificultar la satisfacción de la necesidad de información del usuario aunque la recuperación haya sido eficaz (**Shneiderman**, 1992b; **Baeza-Yates**, 2011; **Hearts**, 2009; **Baeza-Yates**; **Broder**; **Maarek**, 2011).

Aunque el desarrollo científico de la visualización no haya corrido de la mano de otros aspectos incluidos en la recuperación, es posible identificar una serie de modelos estándar y ciertas tendencias.

2. Modelos estándar de presentación de resultados

Los sistemas de recuperación suelen presentar los resultados de una consulta en forma de listado plano unidimensional. Los usuarios, para refinar esos resultados obtenidos, interaccionan con ellos a partir de operaciones de filtrado.

Los criterios más utilizados en la organización de la lista de resultados son: orden, relevancia, recomendación y *clustering* (**Rosenfeld**; **Morville**, 2006; **Pérez-Montoro**, 2010):

- el orden organiza la lista de resultados utilizando como criterio la dimensión alfabética o numérica de alguna de las características (nombre del autor o fecha de creación, por ejemplo) del documento recuperado;
- la relevancia permite organizar en forma de ranking los documentos recuperados utilizando como criterio la adecuación de la consulta del usuario con el contenido del documento;
- la recomendación permite ordenar los resultados por el número de recomendaciones sugeridas por otros usuarios que han consumido previamente ese resultado;
- el *clustering* presenta los resultados agrupados en diferentes subconjuntos formados por documentos que versan sobre un mismo tema y que lo abordan con un enfoque similar (**Larson**, 1991; **Tryon**, 1939).

"La recuperación de información es una estrategia, basada en la interrogación, para la rápida localización de documentos de un fondo que puedan satisfacer las necesidades informativas de un usuario"

Estas formas de organizar los resultados, aunque utilizadas por una parte importante de los sistemas de recuperación, presentan importantes limitaciones:

El criterio de un orden alfabético o numérico no ofrece información extra para que el usuario pueda decidir qué documentos de la lista satisfacen de forma adecuada su necesidad de información temática.

En el caso de la relevancia, el sistema ofrece un ranking colocando en las primeras posiciones aquellos documentos que podrían satisfacer la necesidad temática de un usuario, pero no proporciona información extra sobre el enfoque o la estructura interna del contenido del documento.

El criterio de la recomendación presenta en las primeras posiciones los documentos recomendados por otros usuarios, pero tampoco ofrece información extra sobre el enfoque o la estructura interna del contenido del documento.

Por último, el *clustering* sí ofrece información extra sobre el enfoque del contenido del documento recuperado, pero no orienta al usuario sobre su distribución y estructura temática.

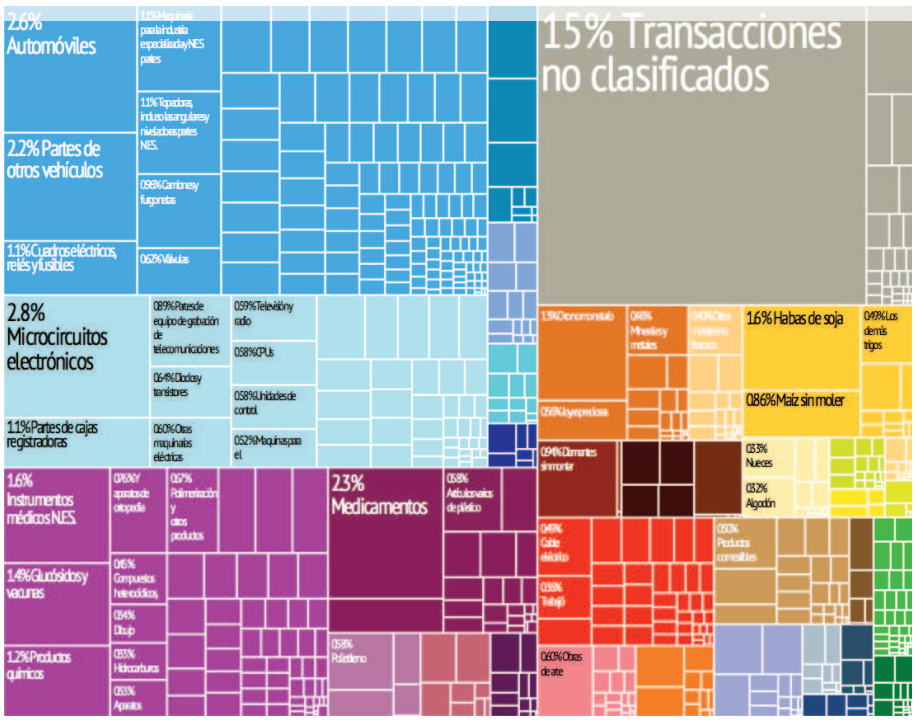

Figura 1. Ejemplo de *treemap*. *http://es.m.wikipedia.org/wiki*

3. Tendencias en la visualización de resultados

En los últimos años, al margen de estos tipos de presentación de resultados, y para superar algunas de sus limitaciones, se han elaborado diferentes propuestas visuales para mejorar la interacción de los usuarios con los resultados recuperados.

Figura 2. Ejemplo de *tag cloud*. *http://www.flickr.com/photos/artangelo/6225813841*

La mayoría de esas propuestas pueden articularse en tres grandes grupos: las visualizaciones de *clustering*, las basadas en la visualización de los términos de la consulta y las que utilizan *thumbnail images* (imágenes miniaturizadas) de los documentos.

Visualizaciones de *clusterings*

Intentan representar las categorías y las relaciones entre éstas bajo las que pueden agruparse los documentos recuperados. Las principales tendencias se basan en la utilización de, entre otros, *treemaps*, *tag clouds* o *network graphs*.

–*Treemaps*

Representan las relaciones jerárquicas de un conjunto de categorías mediante la utilización de rectángulos anidados (*nested*) optimizando el espacio utilizado para la presentación (**Shneiderman**, 1992a; **Shneiderman**; **Plaisant**, 2009). Cada rectángulo tiene unas dimensiones proporcionales a la cantidad de documentos recuperados bajo esa categoría (figura 1). Normalmente los rectángulos están coloreados para favorecer su lectura por parte de los usuarios.

–*Tag clouds* (nubes de etiquetas)

Representan las categorías en forma de nube de palabras, donde el color y el tamaño de la palabra codifican si hay o no muchos documentos recuperados bajo esa categoría (**Begelman**; **Keller**; **Smadja**, 2006). Normalmente las etiquetas que aparecen en la nube suelen ser hipervínculos que llevan al listado de los documentos que han sido recuperados bajo esa etiqueta (figura 2).

–*Network graphs* (gráficos de red)

Representan cada categoría como un objeto y las relaciones entre las categorías mediante líneas o curvas (figura 3). Según la interacción más habitual, si se clica en un objeto o categoría obtendremos el listado de documentos recuperados clasificados bajo esa categoría. Existen muchos ejemplos en los que se ha aplicado este tipo de visualización (**Moya-Anegón** *et al.*, 2004; **Granitzer** *et al.*, 2004, **Brandes**; **Hoefer**; **Lerner**, 2006).

> "Nuevas propuestas de visualización son: *clustering* (*treemaps*, *tag clouds* y *network graphs*), representación de los términos de la consulta y utilización de *thumbnail images*"

Visualización de los términos de la consulta

Las visualizaciones basadas en la representación de los términos de la consulta (*query terms*) propuestos por el usuario acostumbran a utilizar dos estrategias: la visualización de los términos en el propio documento recuperado o en la página de resultados (**Hearst**, 2009).

- En el primer caso, nos presenta el documento recuperado resaltando en el mismo aquellas palabras del texto que coincidan literalmente con los términos de la consulta realizada (**Egan** *et al.*, 1989). Algunos estudios señalan que los usuarios prefieren ver implementada esta técnica mediante el uso de color en el resaltado de las palabras del texto que coinciden con los términos de la consulta (**Hornbæk**; **Frøkjær**, 2001).
- En el segundo caso, en la página de resultados, cada documento es representado como una barra horizontal proporcional a su extensión donde se ubican pequeños cuadrados con los términos de la consulta que aparecen en el texto (**Hoeber**; **Yang**, 2006). Igual que en el caso

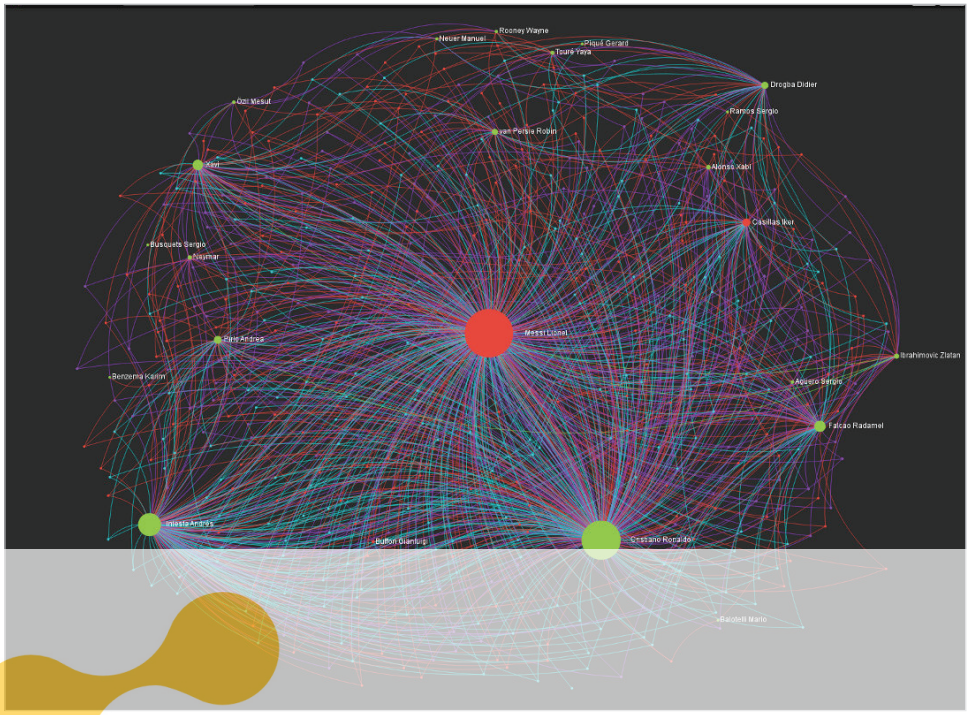

Figura 3. Ejemplo de *network graph*. *http://www.flickr.com/photos/yaph/8552474453*

anterior, algunos estudios señalan que estas visualizaciones mejoran para los usuarios cuando se complementan introduciendo una escala de color en los cuadrados que represente la frecuencia de los términos de la consulta en el texto del documento (**Anderson** *et al.*, 2002).

> **"Los criterios más utilizados en la organización de resultados de una búsqueda son: orden, relevancia, recomendación y *clustering*"**

Thumbnail images

Esta técnica se fundamenta en el hecho de que el sistema visual humano permite capturar los rasgos esenciales de una imagen completa en 110 milisegundos o menos, justo lo que se tarda en leer sólo una o dos palabras (**Woodruff** *et al.*, 2001). Algunos estudios defienden que introducir estas imágenes en los resultados de búsqueda puede funcionar como resúmenes visuales de los documentos para los usuarios (**Jhaveri**; **Räihä**, 2005).

4. Limitaciones en las propuestas visuales

Frente a las listas de resultados más estándares, las nuevas propuestas de visualización descritas pueden mejorar la experiencia de búsqueda de los usuarios en un sistema de recuperación. Sin embargo, presentan también importantes limitaciones.

Respecto a las visualizaciones de *clusterings*:

- los *treemaps* ofrecen información extra sobre el enfoque temático del contenido del documento recuperado y las posibles relaciones semánticas que pueden mantener con otros documentos recuperados, pero no orientan al usuario sobre la distribución y estructura temática de cada uno de esos documentos.
- Los *tag clouds* ofrecen información extra sobre el enfoque temático del contenido del documento recuperado pero no sobre las posibles relaciones semánticas que pueden mantener con otros documentos recuperados ni tampoco orientan al usuario sobre la distribución y estructura temática de cada uno de esos documentos.
- Los *network graphs* ofrecen información extra sobre el enfoque temático del contenido del documento recuperado y las posibles relaciones semánticas que pueden mantener con otros documentos recuperados, pero no orientan al usuario sobre la distribución y estructura temática de cada uno de esos documentos. En este caso se une también el problema de que cuando el *network graph* incluye muchos objetos y relaciones, el usuario no puede explorarlo de una forma cómoda, viéndose obligado a utilizar el zoom para tener una visión global del *network* o explorar parcialmente las áreas de éste que le interesen (**Viégas**; **Donath**, 2004). Algunos autores defienden estrategias parciales para mejorar esta última forma de visualización focalizando la visualización sobre el nodo que le interesa al usuario (**Yee** *et al.*, 2001) o eliminando de la visualización aquellos nodos que no han sido clicados por este (**Fellbaum**, 1998).

Las visualizaciones basadas en la representación de los términos de la consulta (*query terms*) presentan también importantes limitaciones:

- sólo ofrecen documentos en los que aparezcan los términos de consulta. No ofrecen informa-

ción extra sobre el enfoque temático del contenido del documento recuperado y las posibles relaciones semánticas que pueden mantener con otros documentos recuperados;

- no orientan al usuario sobre la distribución y estructura temática no relacionada con esos términos en cada uno de esos documentos recuperados.

La estrategia de completar la lista de resultados con *thumbnail images* de los documentos recuperados también presenta importantes limitaciones. Estas visualizaciones, aunque complementarias, no ofrecen información extra sobre el enfoque temático del contenido del documento recuperado, ni sobre las posibles relaciones semánticas que pueden mantener con otros documentos recuperados, ni orientan al usuario sobre la distribución y estructura temática de cada uno de esos documentos. En esta línea existen estudios que muestran que esta estrategia no mejora significativamente la experiencia de búsqueda de los usuarios (**Czerwinski** *et al.*, 1999; **Dziadosz**; **Chandrasekar**, 2002), aunque pueden servir de ayuda en parte si se agrandan las imágenes (**Kaasten**; **Greenberg**; **Edwards**, 2002)

5. Conclusiones

Como se desprende de este análisis, tanto las propuestas estándar de presentación de resultados como las tendencias visuales en la recuperación ofrecen limitaciones importantes que pueden dificultar la correcta satisfacción de las necesidades informativas por parte de los usuarios.

Sin embargo, es posible establecer una serie de requisitos que sirvan de guía para la mejora de las herramientas de visualización en el proceso de la recuperación. Esos requisitos pueden clasificarse en dos grandes grupos:

- Relacionados con los aspectos arquitectónicos del sistema: una buena herramienta de visualización debe ofrecer al usuario básicamente tres funciones: control sobre el proceso de recuperación, posibilidad de agregación o desagregación de los documentos recuperados (estrechamiento y ampliación de los resultados de búsqueda) y navegabilidad de la página de resultados (para facilitar su exploración).
- Relacionados con las características semánticas de los documentos: una buena herramienta debe comenzar representando cada uno de los documentos recuperados con una adecuada densidad de información asociada. Esa densidad de información debe mantener el equilibrio entre la cantidad mínima de información necesaria para que el usuario pueda identificar y discriminar el contenido del documento y la cantidad de información máxima para que el sistema pueda presentar de forma visual la totalidad del conjunto de documentos recuperados.

Manteniendo ese equilibro en la densidad de información ofrecida por documento, el sistema debe también suministrar información sobre el enfoque temático del contenido del documento recuperado, debe mostrar las posibles relaciones semánticas que éste puede mantener con otros documentos recuperados, y debe también orientar al usuario sobre la distribución y estructura temática de cada uno de esos documentos recuperados.

6. Bibliografía

Anderson, Terry; **Hussam, Ali**; **Plummer, Bill**; **Jacobs, Nathan** (2002). "Pie charts for visualizing query term frequency in search results". En: *Proceedings of the 5th International Conference on Asian Digital Libraries: Digital Libraries: People, Knowledge, and Technology*, pp. 440–451. London: Springer-Verlag.
http://dx.doi.org/10.1007/3-540-36227-4_52

Baeza-Yates, Ricardo (2011). "Tendencias en recuperación de información en la web". *BiD: textos universitaris de biblioteconomia i documentació*, n. 27.
http://www.ub.edu/bid/27/baeza2.htm

Baeza-Yates, Ricardo; **Broder, Andrei**; **Maarek, Yoelle** (2011). "The new frontier of web search technology: seven challenges". En: Ceri, Stefano; Brambilla, Marco (Eds.). *Search Computing* (v. 6585, pp. 3–9). Berlin & Heidelberg: Springer Verlag.
http://dx.doi.org/10.1007/978-3-642-19668-3_1

Baeza-Yates, Ricardo; **Ribeiro-Neto, Berthier** (1999). *Modern information retrieval*. Boston, MA: Addison-Wesley Longman. ISBN: 978 0201398298

Bates, Marcia J. (1989). "The design of browsing and berrypicking techniques for the online search interface". *Online information review*, n. 13, pp. 407-424.
http://dx.doi.org/10.1108/eb024320

Begelman, Grigory; Keller, Philipp; Smadja, Frank. (2006). "Automated tag clustering: Improving search and exploration in the tag space". En: *Collaborative Web Tagging Workshop at WWW2006*, Edinburgh, Scotland, pp. 15–33.
http://www.ra.ethz.ch/cdstore/www2006/www.rawsugar.com/www2006/20.pdf

Brandes, Ulrik; **Hoefer, Martin**; **Lerner, Jürgen** (2006). "WordSpace: visual summary of text corpora". En: Erbacher, Robert F.; Roberts, Jonathan C.; Gröhn, Matti T.; Börner, Katy (Eds.). En: *Proceedings of SPIE. Visualization and data analysis 2006*, v. 6060, pp. 212-223. Bellingham, WA: SPIE-the International Society for Optics and Photonics.
http://www.mpi-inf.mpg.de/~mhoefer/05-07/Brandes06Wordspace.pdf

Czerwinski, Mary P.; **Van-Dantzich, Maarten**; **Robertson, George**; **Hoffman, Hunter** (1999). "The contribution of thumbnail image, mouse-over text and

spatial location memory to web page retrieval in 3D". En: *Proceedings of the INTERACT'99 conference*, pp. 163-170. Dordrecht, Kluwer.
http://research.microsoft.com/en-us/um/people/marycz/interact99.pdf

Dziadosz, Susan; **Chandrasekar, Raman** (2002). "Do thumbnail previews help users make better relevance decisions about web search results?" En: *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 365-366. New York, NY: ACM Press.
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.94.5957&rep=rep1&type=pdf
http://dx.doi.org/10.1145/564376.564446

Egan, Dennis E.; **Remde, Joel R.**; **Gomez, Louis M.**; **Landauer, Thomas K.**; **Eberhardt, Jennifer**; **Lochbaum, Carol C.** (1989). "Formative design evaluation of superbook". *ACM transactions on information systems (TOIS)*, v. 7, n. 1, pp. 30–57.
http://dx.doi.org/10.1145/64789.64790

Fellbaum, Christiane (1998). *WordNet: an electronic lexical database*. Massachusetts: MIT Press. ISBN: 978 0262061971

Granitzer, Michael; **Kienreich, Wolfgang**; **Sabol, Vedran**; **Andrews, Keith**; **Klieber, Werner** (2004). "Evaluating a system for interactive exploration of large, hierarchically structured document repositories". En: *IEEE Symposium on Information Visualization, 2004. INFOVIS 2004*, pp. 127-134. Los Alamitos, CA: IEEE Computer Society Press.
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.123.2297&rep=rep1&type=pdf
http://dx.doi.org/10.1109/INFVIS.2004.19

Hearts, Marti (2009). *Search user interfaces*. Cambridge: Cambridge University Press. ISBN: 978 0521113793

Hoeber, Orland; **Yang, Xue-Dong** (2006). "A comparative user study of web search interfaces: HotMap, concept highlighter, and *Google*". En: *WI'06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 866-874. Washington, DC: IEEE Computer Society.
http://dx.doi.org/10.1109/WI.2006.6

Hornbæk, Kasper; **Frøkjaer, Erik** (2001). "Reading of electronic documents: the usability of linear, fisheye, and overview+ detail interfaces". En: *CHI'01 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 293-300. New York, NY: ACM Press.
http://dx.doi.org/10.1145/365024.365118

Jhaveri, Natalie; **Räihä, Kari-Jouko** (2005). "The advantages of a cross-session web workspace". En: *CHI'05 Extended Abstracts on Human Factors in Computing Systems*, pp. 1.949-1.952. New York, NY: ACM Press.
http://www.cs.uta.fi/~nj73504/jhaveri_chi05.pdf
http://dx.doi.org/10.1145/1056808.1057064

Kaasten, Shaun; **Greenberg, Saul**; **Edwards, Christopher** (2002). "How people recognize previously seen Web pages from titles, URLs and thumbnails". En: Faulkner, Xristine; Finlay, Janet; Détienne, Françoise. *People and Computers XVI - Memorable Yet Invisible: Proceedings of HCI 2002*, pp. 247–266. Berlin/Heidelberg: Springer.
http://dx.doi.org/10.1007/978-1-4471-0105-5_15

Larson, Ray R. (1991). "Classification clustering, probabilistic information retrieval, and the online catalog". *The library quarterly*, v. 61, n. 2, pp. 133–173.

Moya-Anegón, Félix; **Vargas-Quesada, Benjamín**; **Herrero-Solana, Víctor**; **Chinchilla-Rodríguez, Zaida**; **Corera-Álvarez, Elena**; **Munoz-Fernández, Francisco-José** (2004). "A new technique for building maps of large scientific domains based on the cocitation of classes and categories". *Scientometrics*, v. 61, n. 1, pp. 129–145.
http://eprints.rclis.org/10066

Pérez-Montoro, Mario (2010). "Arquitectura de la información en entornos web". *El profesional de la información*, v. 19, n. 4, pp. 333-337.
http://diposit.ub.edu/dspace/handle/2445/23507
http://dx.doi.org/10.3145/epi.2010.jul.01

Rosenfeld, Louis; **Morville, Peter** (2006). *Information architecture for the world wide web: designing large-scale web sites*. Sebastopol, CA: O'Reilly Media.

Shneiderman, Ben (1992a). *Designing the user interface: strategies for effective human-computer interaction*. (2nd ed.) Boston, MA: Addison-Wesley Longman Publishing Co., Inc. ISBN: 978 0321537355

Shneiderman, Ben (1992b). "Tree visualization with tree-maps: 2-d space-filling approach". *ACM transactions on graphics*, v. 11, n. 1, pp. 92–99.
https://www.cs.umd.edu/users/ben/papers/Shneiderman1992Tree.pdf

Shneiderman, Ben; **Plaisant, Catherine** (2009). *Twenty years of treemaps: a personal visualization of hierarchies.*
http://www.cs.umd.edu/hcil/treemap-history

Tryon, Robert (1939). *Cluster analysis*. New York, NY: McGraw-Hill. ISBN: 978 0226813127

Viégas, Fernanda B.; **Donath, Judith** (2004). "Social network visualization: can we go beyond the graph". *Workshop on Social Networks, CSCW 2004*, pp. 6-10.
http://alumni.media.mit.edu/~fviegas/papers/viegas-cscw04.pdf

Woodruff, Allison; **Faulring, Andrew**; **Rosenholtz, Ruth**; **Morrison, Julie**; **Pirolli, Peter** (2001). "Using thumbnails to search the Web". En: *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 198–205. New York, NY: ACM Press.
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.28.3774&rep=rep1&type=pdf
http://dx.doi.org/10.1145/365024.365098

Yee, Ka-Ping; **Fisher, Danyel**; **Dhamija, Rachna**; **Hearst, Marti** (2001). "Animated exploration of dynamic graphs with radial layout". En: *INFOVIS '01 Proceedings of the IEEE Symposium on Information Visualization 2001*, (p. 43). Washington, DC: IEEE Computer Society. ISBN: 0 7695 1342 5